Resource Allocation in Virtual Desktop Clouds: VMLab-GENI Experiment

Project Team: **Prasad Calyam,** Ph.D. pcalyam@osc.edu,
Aishwarya Venkataraman, Mukundan Sridharan, Ph.D.,
Alex Berryman, Rohit Patali

Research Sponsors: NSF CNS-1050225, Dell, VMware, IBM

GEC11 Experimenter Lightning Talk
July 27 2011

Topics of Discussion

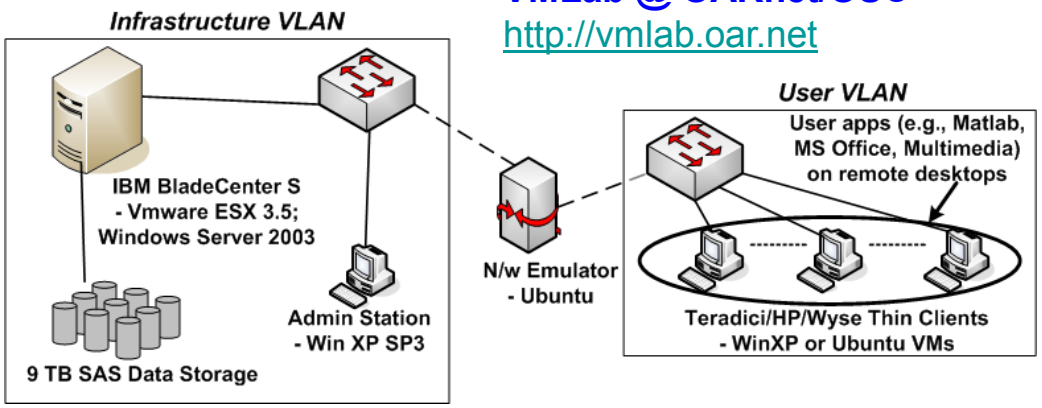
- VMLab-GENI Experiment Context
- VDC Research Problem and Solution
- GEC10 Experiment Demonstration
- New Experiments Planned

Topics of Discussion

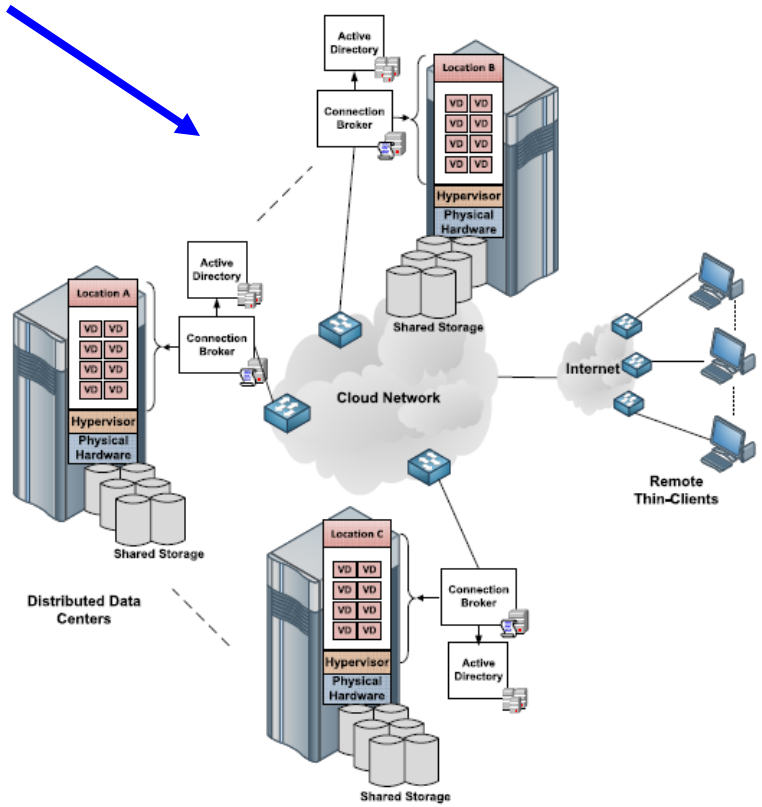
- VMLab-GENI Experiment Context
- VDC Research Problem and Solution
- GEC10 Experiment Demonstration
- New Experiments Planned

VMLab-GENI Experiment Context

VMLab @ OARnet/OSC
http://vmlab.oar.net



- *VMLab → GENI*
  - *Measurement & Evaluation →*
    Allocation & Management
  - *Tabletop Experiments →*
    Cloud Experiments



Virtual Desktop Cloud (VDC)
in the VMLab-GENI Infrastructure

Motivations for Research

- Recent advances in thin clients and the numerous benefits in transitioning user desktops to virtual desktop clouds (VDCs)
  - Convenience, Cost savings, Green IT, Security, …

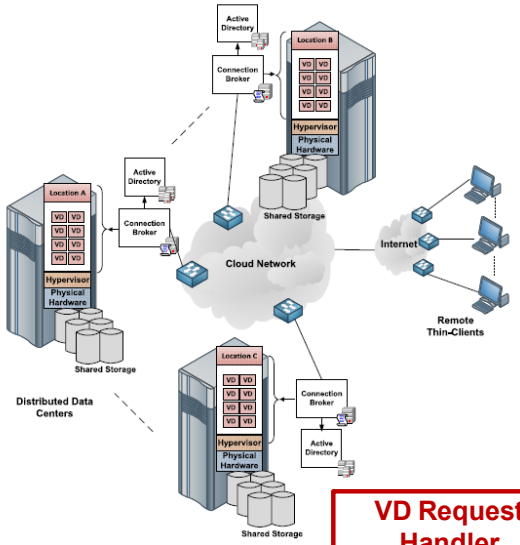
- *Need for "system-aware", "network-aware", "human-aware" frameworks and tools to deploy virtual desktop clouds*
  - Existing work focuses mainly upon system (i.e., CPU and memory) measurements for server-side resource adaptation
  - Our focus is to couple client-and-server resource adaptation with measurements of network health and user experience
    - Minimize cloud resource over-commitment
    - Avoid guesswork in configuring thin client protocols
    - Deliver optimum user experience of virtual applications

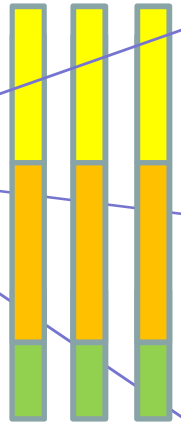
VDCs Today – Overprovisioning and Guesswork…

- High consistent CPU
- High consistent memory
- High bandwidth connectivity

Research Scientist

VDC Service Provider





VD Request Handler

- Low bursty CPU
- Low bursty memory
- Medium bandwidth connectivity

Home User

- Low bursty CPU
- Low bursty memory
- Low bandwidth connectivity

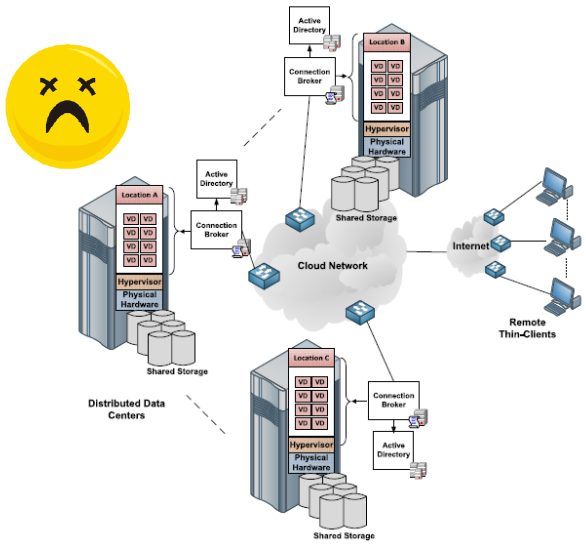
Mobile User

Fixed Resource Allocation Model = Available Resources / Number of Users

Overprovisioning and Guesswork Fails!

- Calls from unhappy customers
- High operation $$

VDC Service Provider



- Inadequate CPU, memory and bandwidth

(Impact e.g., Slow interaction response times)

Research Scientist

- Inadequate CPU, memory and bandwidth

(Impact e.g., IPTV with impairments and slow playback)

Home User

- Excess CPU, memory and bandwidth

(Impact e.g., Good interaction response times and smooth IPTV playback)

Mobile User

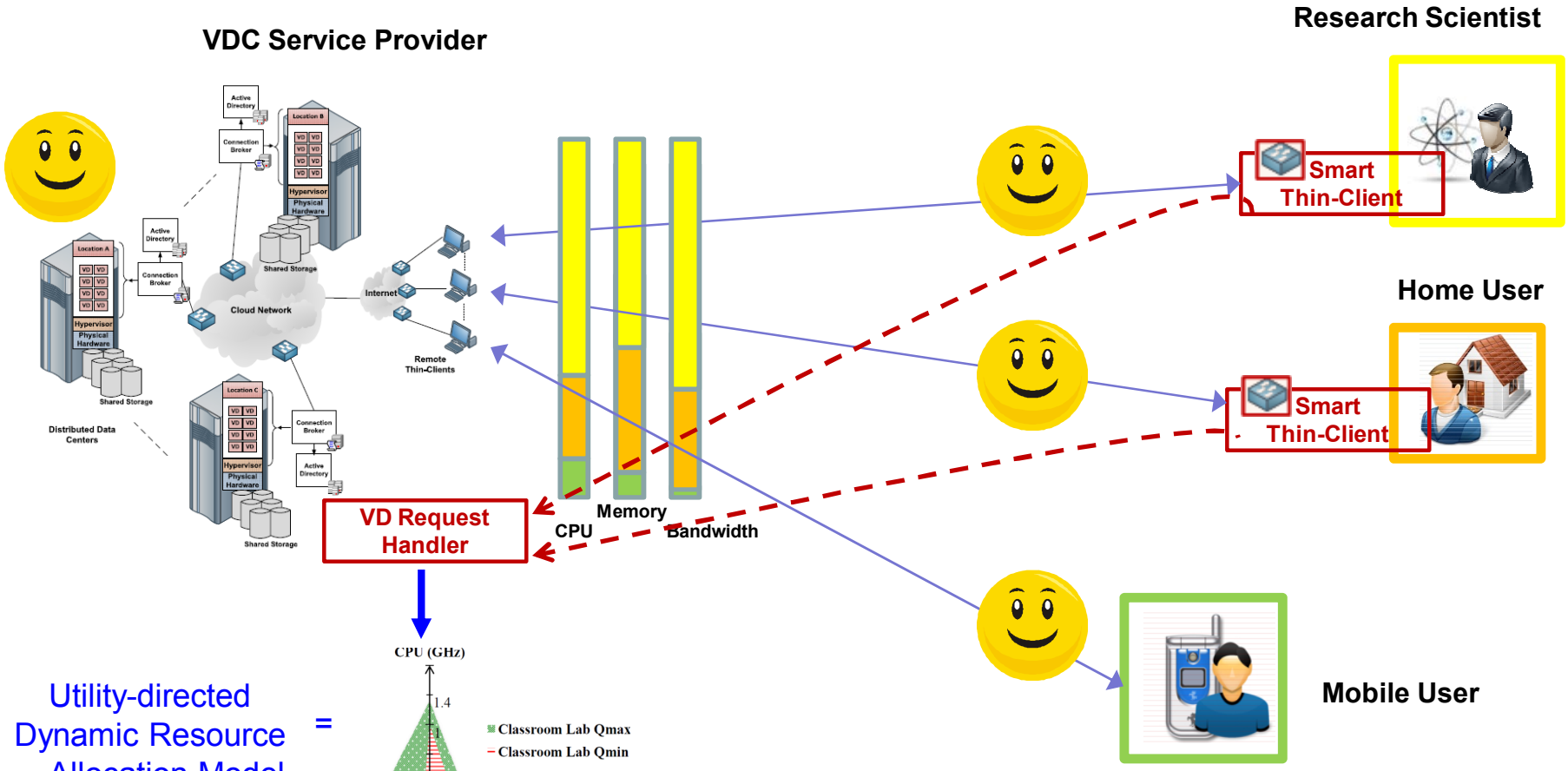
Problem: Resource allocation without awareness of system, network and user experience characteristics

VDCs in the Future – Smart set-top boxes at user sites

- Happy customers
- Low operation $$

- Utility-directed CPU, memory and bandwidth

(Impact e.g., Good interaction response times and smooth IPTV playback)



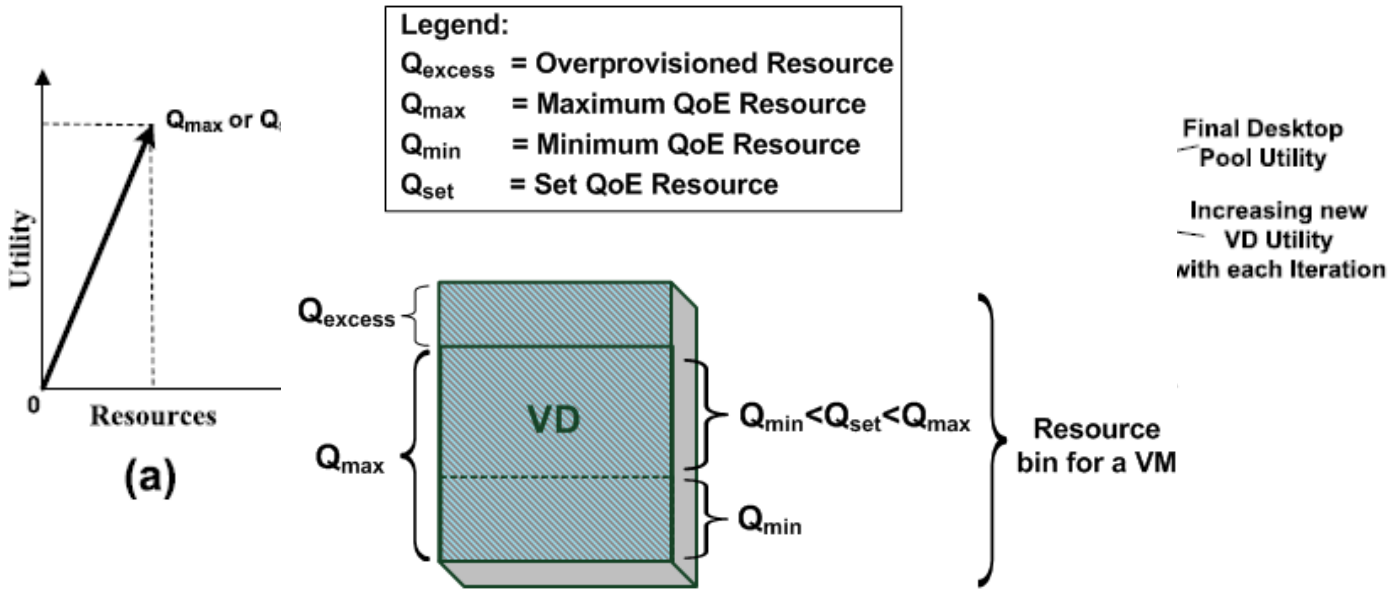
Topics of Discussion

- VMLab-GENI Experiment Context
- VDC Research Problem and Solution
- GEC10 Experiment Demonstration
- New Experiments Planned

Utility-directed Resource Allocation Model (U-RAM)

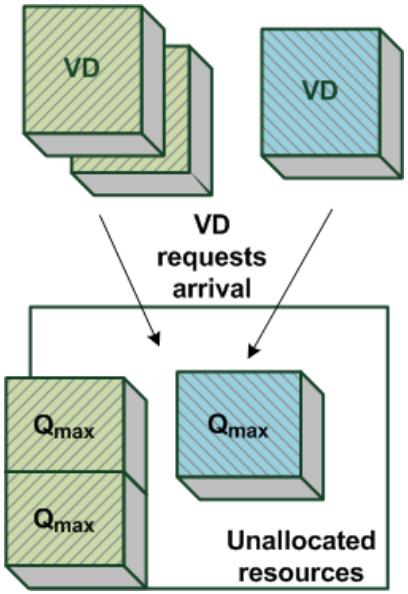
(Published in Computer Networks (Elsevier) SI on Internet-based Content Delivery, 2011)

- Utility function indicates how much of application performance in a VD can be increased with larger resource allocation
- Fixed RAM (**F-RAM**) tends to allocate resources that result in Q_{excess}
- **U-RAM** profiles users based on VDBench measurements and allocates resources that results in either Q_{min} /Q_{set} /Q_{max}
- We have developed a novel iterative algorithm for resource allocation that has fast convergence

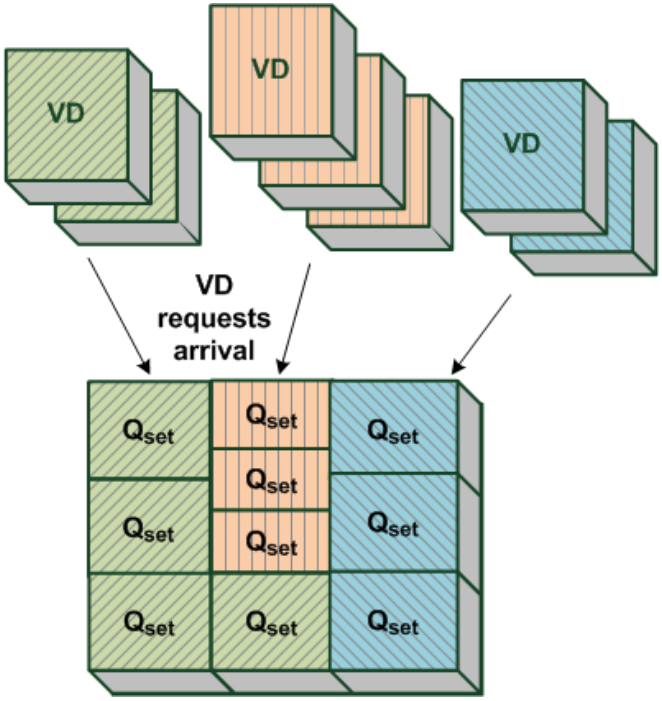


III. U-RAM Iterative Resource Allocation for Desktop utility

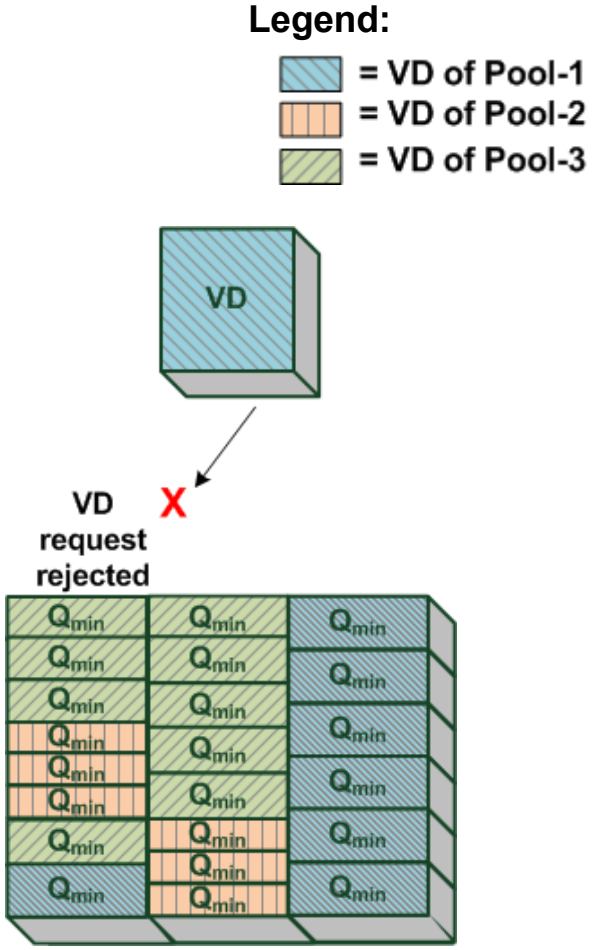
U-RAM Illustration



I. New VD requests handling with freely available resources



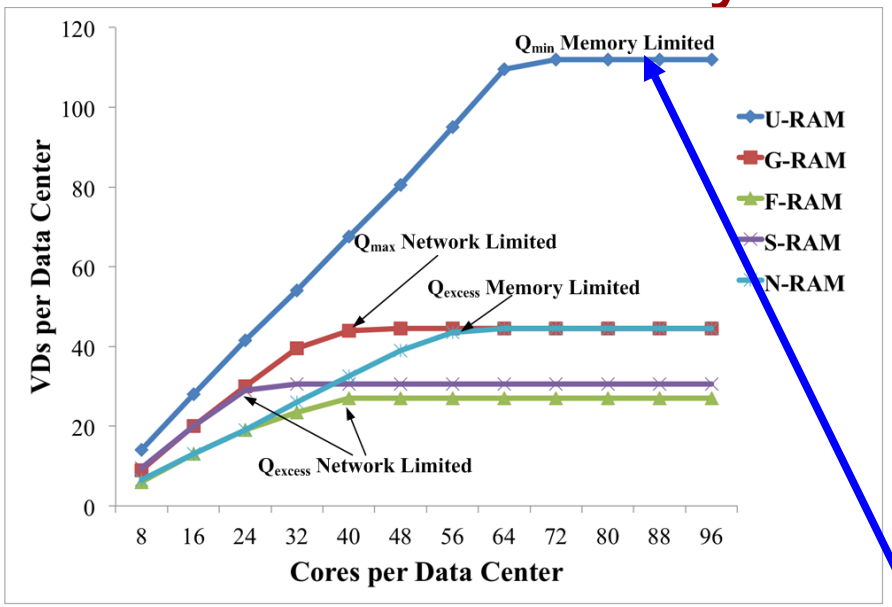
II. New VD requests handling with all available resources allocated

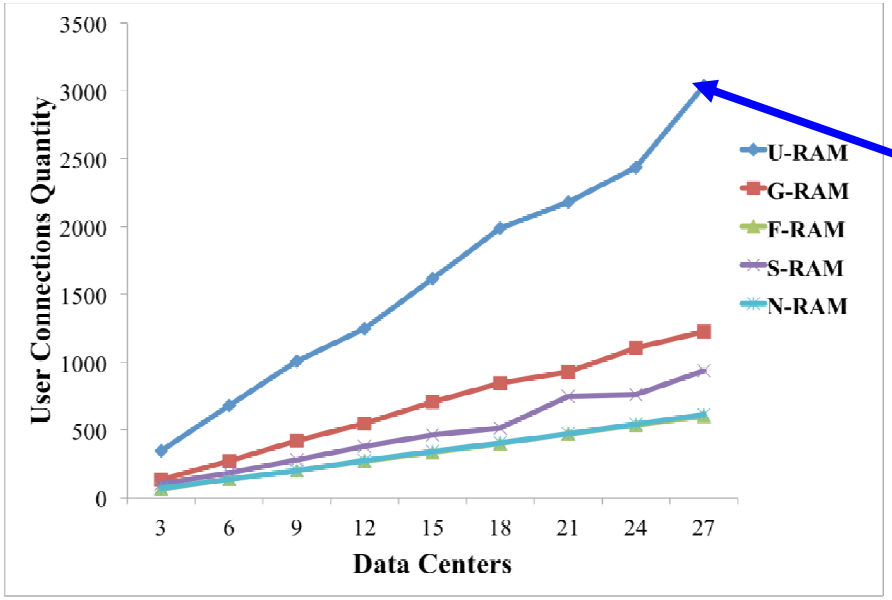


III. New VD request rejected when SLA violation situation occurs

- Kuhn-Tucker Theorem suggests that global utility is maximized if resource allocation is identical to each request
  - VDBench measurements show vast differences in utility functions for subsets of applications in a user group, hence we apply Kuhn-Tucker optimality in U-RAM across “desktop pools”

Cloud Scalability Performance Comparison





- *Fixed RAM* (**F-RAM**): each VD is over provisioned and is given resources that produce utility in Q_{excess} range
- *Network-aware RAM* (**N-RAM**): Allocation is aware of the Q_{max} required for network resources, but over provisions Q_{excess} for system (RAM and CPU) resources due to lack of system awareness information
- *System-aware RAM* (**S-RAM**): Allocation is opposite of N-RAM; Q_{max} is provisioned for the system resources and Q_{excess} is provisioned for the network resources
- *Greedy RAM* (**G-RAM**): Allocation is aware of the Q_{max} requirement in terms of both the system as well as the network resources based purely upon rule-of-thumb information
- *Utility-directed RAM* (**U-RAM**): Allocation operates a VD with utility in:
  - Q_{max} range while there are abundant resources available
  - Q_{set} range when resources are already allocated under low VD request loads
  - Q_{min} range when resources are already allocated under high VD request loads

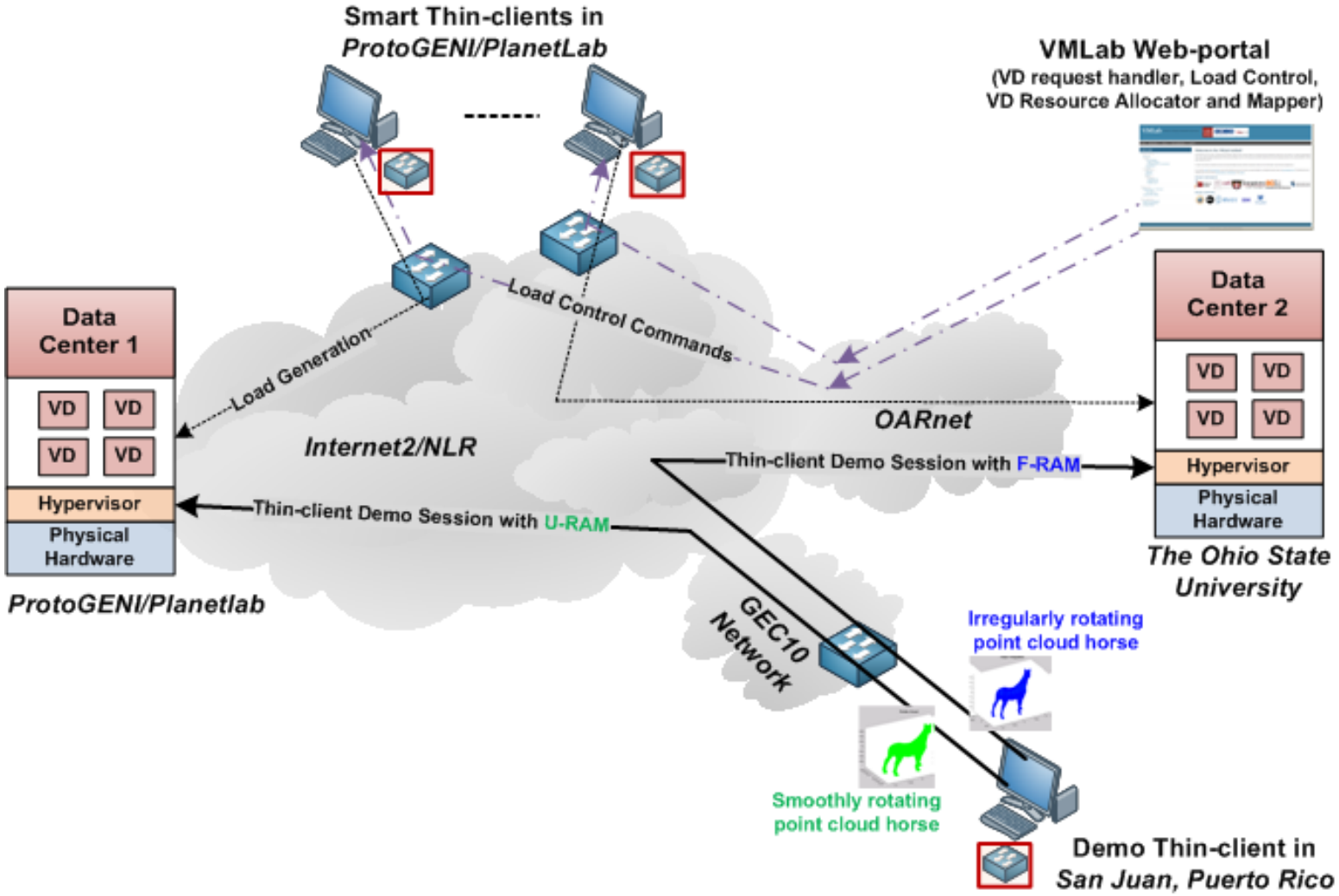
Topics of Discussion

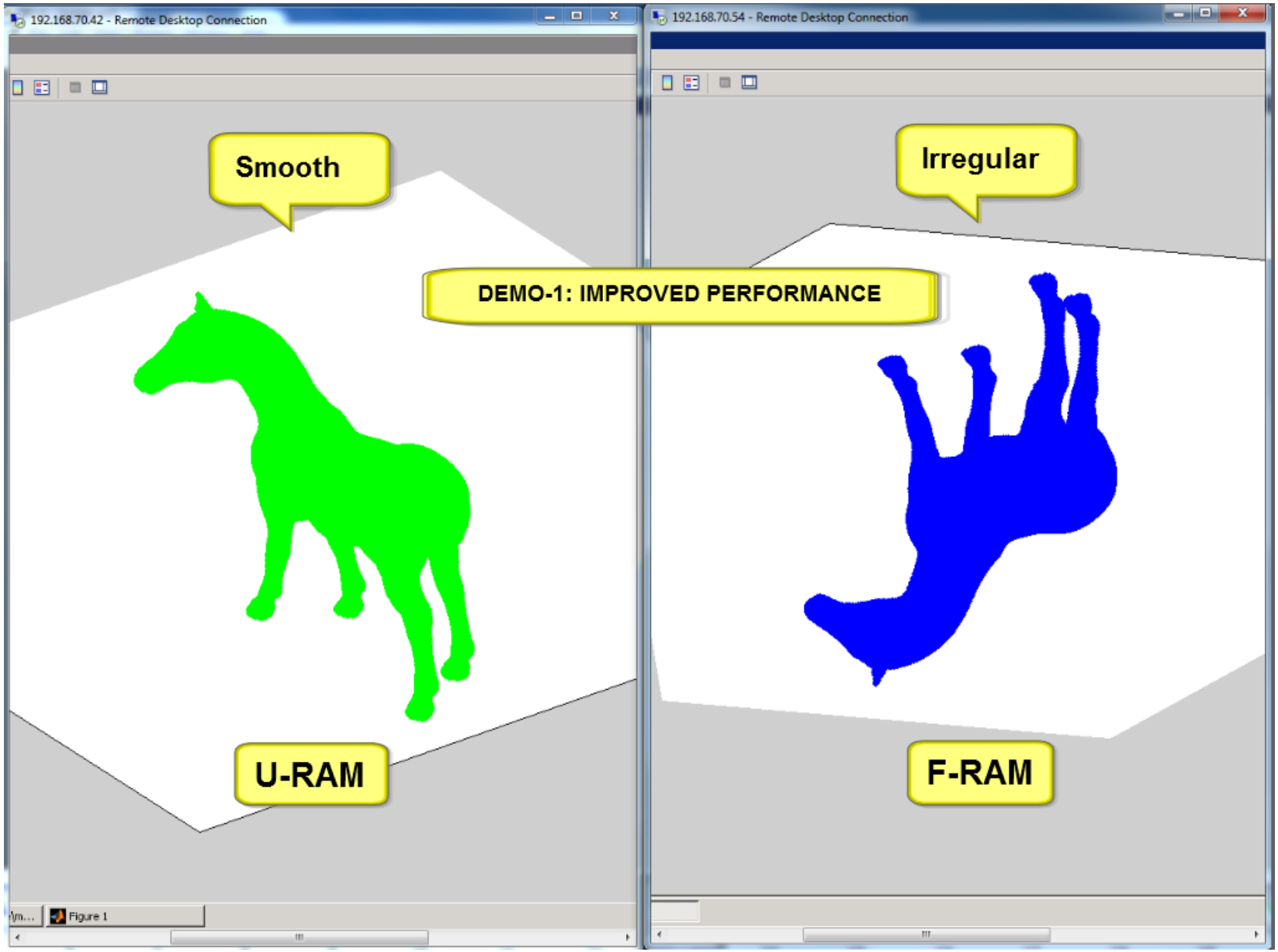
- VMLab-GENI Experiment Context
- VDC Research Problem and Solution
- GEC10 Experiment Demonstration
- New Experiments Planned

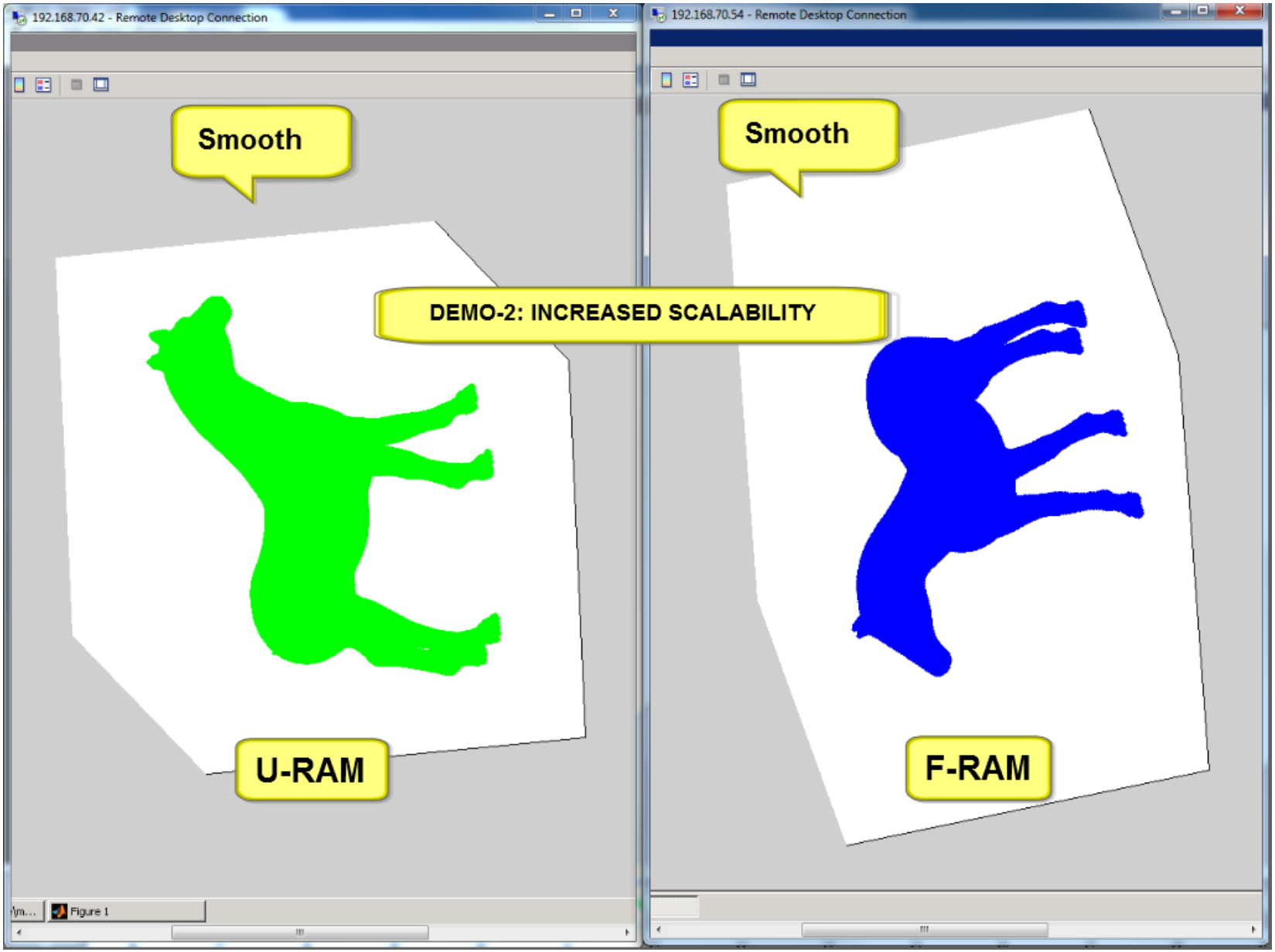
GEC10 Experiment Demonstration

- Compared U-RAM and F-RAM performance
- Created a datacenter in VMLab
  - One physical server each in VMLab for U-RAM and F-RAM
- VDC clients on ProtoGENI slice in GEC10
  - Developed a web-portal to launch VDC clients and control network emulation for demonstration
- Leveraged OnTimeMeasure's new metric creation capability
  - Path-based measurements of network health such as delay, available bandwidth, loss
  - Host-based measurements from VMware VDI tools such as CPU, memory, number of VM connections

GEC10 Experiment Demonstration Setup




192.168.70.42 - Remote Desktop Connection
192.168.70.54 - Remote Desktop Connection
Smooth
Irregular
DEMO-1: IMPROVED PERFORMANCE
U-RAM
F-RAM
Figure 1


192.168.70.42 - Remote Desktop Connection
Smooth
U-RAM
Figure 1
192.168.70.54 - Remote Desktop Connection
Smooth
F-RAM
DEMO-2: INCREASED SCALABILITY

Topics of Discussion

- VMLab-GENI Experiment Context
- VDC Research Problem and Solution
- GEC10 Experiment Demonstration
- New Experiments Planned

New Experiments Planned

- Slices with distributed datacenters and users
  - Extend GEC10 experiment to a virtual desktop cloud with 3 data centers with L2/L3 network connectivity
  - Leverage GENI capabilities for Experimenters
    - ProtoGENI, OnTimeMeasure, Gush, INSTOOLS, OpenFlow, …
- Improve our Provisioning and Placement Algorithms
  - Develop schemes to account for cost-benefit analysis and cost-of-failure in VDC resource allocations
  - Gather user workload models from task profiling
    - University of Alaska, Fairbanks is collaborating with us on obtaining user workload profiles from NSF funded RAVE project classroom labs
  - Enhance VDBench thin-client embedded software for user and network performance characterization
- Bring actual users into GENI VDC Experiment
  - Fits well with the US IGNITE program to reach city user communities

Thank you for your attention!